More elaborate functions for automatically identifying discriminating features in sets of structures are being developed. Currently, these experimental routines (contained within the "PLAN" program) can be used to analyze functionality, or to identify differences in the ways that superatoms have been imbedded in structures. These routines will shortly be capable of exploiting a simplified library of chemical/spectral tests for particular substructural features; this will allow the program to identify possible discriminating experiments. The current capabilities of these functions are described in subsequent sections.

2.3.1 EXAMINE

The EXAMINE function allows for the identification and selection of structures characterized by particular combinations of substructures, ring-systems and Isoprene-patterns. Further, if relative merits can be associated with the substructural features, then these merit values can be used to rank the structures. In addition to providing information on the frequency of different structural features, the EXAMINE function allows structures with unacceptable combinations of features to be pruned away.

EXAMINE thus extends both the earlier SURVEY function (which EXAMINE has now totally subsumed) and the PRUNE function in CONGEN. (PRUNE remains in CONGEN because of its greater efficiency in simply rejecting undesired structures.) EXAMINE allows structures to be segregated on the basis of combinations of (desired or undesired) structural features. For example, EXAMINE can be used to segregate structures which possess feature A or B, or generally, any arbitrary Boolean expression of relationships among structural features.

The EXAMINE function involves the following steps:

1) the definition of relevant substructural features.

2) [EXAMINE matches the features to the structures produced by an earlier GENERATE or IMBED step, and summarizes their frequency.]

3) [if some form of merit rating is being used, then details of the ranking process are provided.]

4) then, in "EXAMINE sub-command" mode, subsets of structures possessing different combinations of features may be selected. Features may be combined using standard AND/OR/XOR/NOT operators. These subset selection procedures are basically non-destructive; however, it is possible to use them to prune the structure list.

5) if examination of the structures has suggested additional selection features, then the entire process may be repeated (information on the current selection features being preserved to allow new selection criteria to be combined with those already in existence). Previously defined libraries of selection features can be used, either alone or as a supplement to selection features specified for a particular problem. It is also possible to save the current set of selection criteria for future use.

2.3.1.1 Example - Unknown Metabolite from Human Urine

Use EXAMINE to determine which members of a set of candidate structures possess naturally occurring, alpha-amino acid part structures. The compound for which CONGEN provided structural candidates was an unknown component of human urine. The empirical formula was $C_{15}H_{19}NO_5$. There were 78 structural candidates based on this empirical formula and chemical constraints. Ten of the 78 formally possess an alpha-amino acid substructure (-NHCHCOO-). Examination of these structures proceeded as follows (note that the examination would yield the same results if the entire 78 were examined).

```
 EXAMINE
Do you require simply to prune your structure list?:
Do you want to rank your structures?(Y for Yes, ? for
   explanation):
Do you want to use a library?Y
FILE NAME:AMINOACID.LIBRARY;8 [Old version]
READING <SMITH>AMINOACID.LIBRARY;8
Do you want all substructures in the file?:Y

      (file read OK)
Do you want to enter new selection features?:
ALA-1-? Substructure ALA min/max (1 . ANY) present in 1
   structures.
GLY-1-? Substructure GLY min/max (1 . ANY) present in 0
   structures.
VAL-1-? Substructure VAL min/max (1 . ANY) present in 0
   structures.
LEU-1-? Substructure LEU min/max (1 . ANY) present in 0
   structures.
ILEU-1-? Substructure ILEU min/max (1 . ANY) present in 0
   structures.
THRE-1-? Substructure THRE min/max (1 . ANY) present in 0
   structures.
PHE-1-? Substructure PHE min/max (1 . ANY) present in 2
   structures.
TYR-1-? Substructure TYR min/max (1 . ANY) present in 0
   structures.
PRO-1-? Substructure PRO min/max (1 . ANY) present in 0
   structures.
OH-PRO-1-? Substructure OH-PRO min/max (1 . ANY) present in 0
   structures.
```

```
ASP-1-? Substructure ASP min/max (1 . ANY) present in 1
   structures.
GLU-1-? Substructure GLU min/max (1 . ANY) present in 1
   structures.
BETA-ALA-1-? Substructure BETA-ALA min/max (1 . ANY) present
   in 0 structures.
SER-1-? Substructure SER min/max (1 . ANY) present in 0
   structures.
```

[note that only four of the amino acids have their part structures (-NHCHR-COO-) represented in the set of candidates, alanine (ALA), phenylalanine (PHE), glutamine (GLU) and asparagine (ASP)]

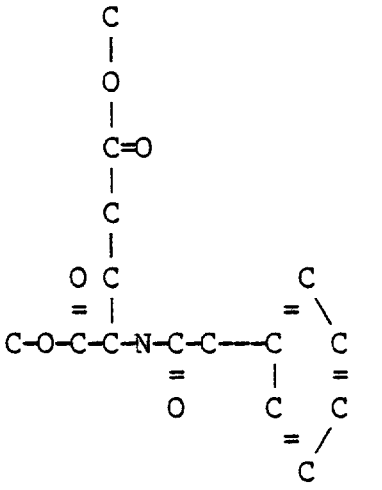
```
Enter commands for selecting subsets of structures with
   particular features.
Do you want help?:
10 STRUCTURES
->SELECT
>(ALA-1-? OR PHE-1-? OR ASP-1-? OR GLU-1-?)

    ...
5 STRUCTURES WITH ((ALA-1-? OR PHE-1-? OR ASP-1-? OR GLU-1-?))
```

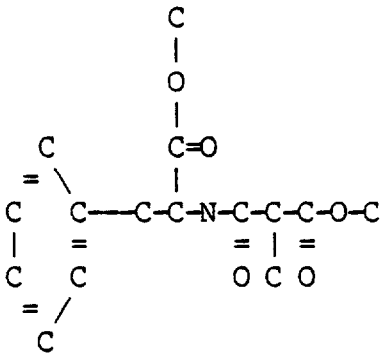
[Only five of the ten (or 78) have any one of the four amino acid substructures. They are drawn below. The first structure drawn is the 77th of the 78 original candidates. The second number refers to its rank based on a comparison of the mass spectrum predicted for this compound against that observed for the unknown. This compound was among the three top-ranked structures (MSRANK) in the original set of 78. It is clearly ranked higher than the other four candidates under the (biochemical) constraint that the compound contain the substructure of a naturally occurring amino acid. Subsequent synthesis and comparison of GC and MS confirmed the identity of the unknown as phenylacetylglutamic acid dimethyl ester.]

->DRAW

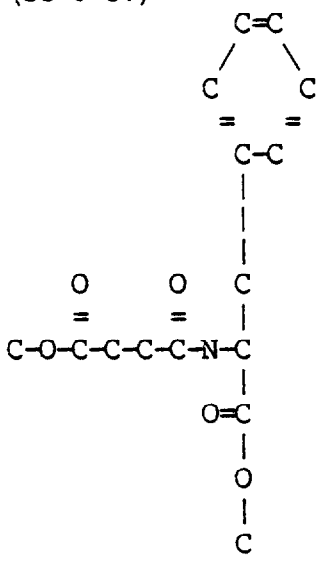
(77 . 84)



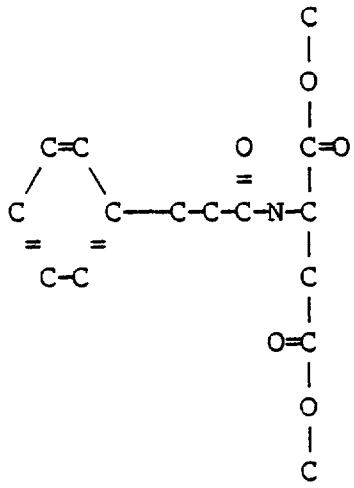
(57 . 57)



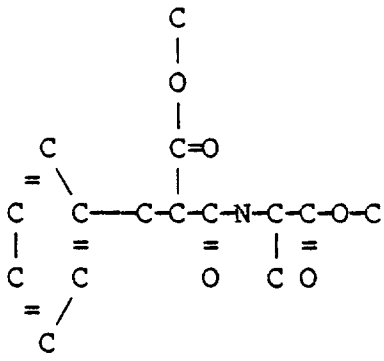
(55 . 57)



(51 . 57)



(29 . 57)



->DONE

2.3.2 PLAN

As mentioned previously, the PLAN program represents our initial efforts toward assembling the heart of an experiment planning program. The goal of PLAN is to identify all structural features which distinguish among structural candidates for an unknown. In the next year we will develop the program which will use this information to suggest experiments. The EXAMINE function, described above, can only look for structural features explicitly supplied by the chemist. Although a summary of such features is quite useful, EXAMINE is insufficient to solve the more general problem of identifying distinguishing substructures.

PLAN in its current form provides the following capabilities:

1) Using a starting substructure supplied by the chemist (for example, one of the superatoms used to construct structural candidates), PLAN can search the local environment of the substructure for distinguishing features, continuing the search until discriminatory characteristics are found.

2) PLAN checks (if requested) for simple differences in the distribution of carbon and hydrogen atoms which could be detected by $^{13}C_{MR}$ or $^{1}H_{MR}$.

3) PLAN can begin at existing functional groups and examine larger substructures by expanding the local environment (as in (1), above) until distinguishing features are found. The example below represents PLAN operated in this mode.

4) PLAN, if requested, performs the operations specified in (3) beginning with double bond systems in the candidates.

2.3.2.1 Example

In the following example, 88 structural candidates for the compound palustrol[8], based on spectroscopic information, were processed by PLAN. The following is a recording of that terminal session. Bracketed comments ([ ]) are inserted to explain the flow of the program.

```
@congen                                              [begin CONGEN]
(<SMITH>CONGEN.;22 . <LISP>CARHART.SAV;70702)
:OK
(LISP)
DO YOU WANT TO SPECIFY AN EMPIRICAL FORMULA?(Y FOR YES):
 RE                                    [RESTORE file of structures]
INPUT FILE:PAL.REACT [Old version]
READING <SMITH>PAL.REACT;2
 THIS IS A FILE WRITTEN BY CONGEN
(COMPOSITION RESTORED)
(EMPIRICAL FORMULA RESTORED)
(AROMATICS RESTORED)
(CONSTRAINTS RESTORED)
USERATOMS  HEP  A1  B1  CH3  CH2  CH  ETH  MET  C  N  O
ALL RESTORED
(88 STRUCTURES)                                      [88 candidates]
 LISP

(LISP):
:(PLAN)                                              [Begin PLAN]

Do you want to specify starting superatoms?     [No starting
                                                 point specified]
Do you want  the program to check  for simple differences  in the
off-resonance decoupled 13c spectrum?Y
```

These structures show no simple differences in their carbon distributions.

Do you want the program to check for simple differences in proton distributions? Y

These structures show no simple differences in their hydrogen distributions.

Do you want the program to check functional groups?Y (See mode (3), above)

Only one substructural class was generated

All compounds have this feature:

```
OH-C
                                    [All compounds possess a
                                    tertiary hydroxyl group,
                                    so PLAN continues]
OH-C
```

present in 88 structures

Only one substructural class was generated

All compounds have this feature:

```
  C
  |
C-C-O                [All compounds have three carbon atoms
  |                  bonded to the tertiary OH, but the
  C                  hydrogen distributions on those carbons
                     differ]
```

By considering proton distributions, 3 subclasses can be distinguished. Do you want to see the protonated structures?Y

```
  CH
C |
H-C-OH
2 |
  CH
```

```
  CH
C |
H-C-OH
2 |
 CH2
```

```
 CH2
C |
H-C-OH
2 |
 CH2
```

[This fact alone is sufficient to consider a dehydration experiment, which is the experiment performed by the chemist when the work was originally done.]

[If desired, each of the three subclasses can be expanded

in turn. Only the expansion of the first class is shown (this class contains the correct structure).]

Do you want this feature to be further enlarged?Y (each subclass will be enlarged separately)

```
  CH
C |
H-C-OH
2 |
  CH
```

[PLAN can continue expansion of each subclass to search for further discriminatory features if requested. The results are omitted for brevity.]

present in 24 structures

```
  CH
C |
H-C-OH
2 |
 CH2
```

present in 49 structures

(end of report)

```
 CH2
C |
H-C-OH
2 |
 CH2
```

present in 15 structures

(end of report)
(continuing now with earlier report stage)
(end of report)
(continuing now with earlier report stage)
(end of report)
Do you want the program to check double bond systems?N

2.4 The Reaction Chemistry Program

During the past year we have made good progress in developing the reaction chemistry program, REACT, into a working tool for laboratory chemists. Two main areas of application are discussed in the subsequent sections. These areas and the examples included are currently in the process of appearing in the literature. Additional details can be obtained by referring to those papers when they appear. The first area of application (subsequent section) is the subject of a paper to appear soon in Tetrahedron. The second area is being written up for publication in the Journal of Chemical Information and Computer Science.

2.4.1 Studies in the Biosynthesis of Natural Products

Manual elucidation of structures arising from chemical reactions which may yield a large number of products via a number of complex, interrelated pathways is a difficult problem. Such reactions are, however, natural candidates for computer-assisted studies because the computer can easily record all intermediates and products as well as interrelationships among them. [22] [1,2] Examples of these reactions include carbonium ion rearrangements, reactions of free radicals and biochemical processes.

REACT is designed to carry out representations of chemical reactions on representations of chemical structures. Reactions, defined by the chemist using the program, are carried out in the synthetic direction as opposed to the retro-synthetic direction of programs for computer-aided synthesis. [3] In structure elucidation problems, the set of structures undergoing reaction is the current set of candidate structures for an unknown. It is clear, however, that the program can also be used effectively in following reactions of a single, known compound participating in a complex sequence of reactions. For example, we showed [22] that CONGEN together with REACT provides a convenient method for studying acid catalyzed rearrangements such as the conversion of tetrahydrodicyclopentadiene to adamantane. In that example, the complete set of isomers was generated by CONGEN. Subsequently, a one-step reaction carried out on each isomer afforded the complete rearrangement graph. An alternative method, similar to that discussed in subsequent sections, is to use a single isomer as a precursor. In the examples given in this work, a single precursor was subjected to repetitive application of a set of reactions.

---

[1] T. M. Gund, P. v. R. Schleyer, P. H. Gund and W. T. Wipke, J.Am.Chem.Soc. 97, 743 (1975).

[2] S. A. Godleski, P. v. R. Schleyer, E. Osawa, Y. Inamoto and Y. Fujikura, J.Org.Chem. 41, 2596 (1976).

[3] E.J. Corey and W.T. Wipke, **Science 166**, 178 (1969).

To demonstrate the utility of REACT we present two examples where a given precursor of known structure is subjected to an extended sequence of reactions. At each step in the sequence one or more reactions may apply to the products from the previous step. As will be shown in the sequel such an approach is especially well suited to problems involving the biosynthesis of natural products. A complete description of this work will appear shortly [22].

2.4.1.1 Generation of Biosynthetically Plausible Sterol Side Chains

Sterols are naturally occurring steroidal alcohols (usually 3-ols) which differ in the number and the position of methyl groups and the degree of unsaturation (present as a double bond or cyclopropyl ring). New sterols are frequently isolated in minute quantities from natural sources. Because of their structural similarities and the large number of different sterols present as a mixture in the same source (a recent paper [4] documents the isolation of ca. 50 sterols from one marine source) it is often difficult to separate them and to obtain pure compounds in quantities large enough for structure determination by conventional methods. Some structural assignments are based on biogenetic considerations, assuming that compounds from the same origin are related to each other through formation along the same biochemical pathway. This pathway can be a series of complicated chemical reactions which yield a large number of intermediates and products. It is difficult to follow manually such a series of reactions in order to explore all possible structural alternatives. To date, over 100 different 3-hydroxy sterols have been isolated, the majority of them based on the seven nuclear skeletons [5].

We use a method of combined gas chromatography/mass spectrometry (GC/MS) to analyze complex mixtures of sterols in a search for new compounds which may represent important biosynthetic intermediates. Part of this method involves research in interpretation and prediction of mass spectra. [23] We have used the REACT program as an additional tool to predict plausible structural candidates to guide both our manual and computer-based interpretations.

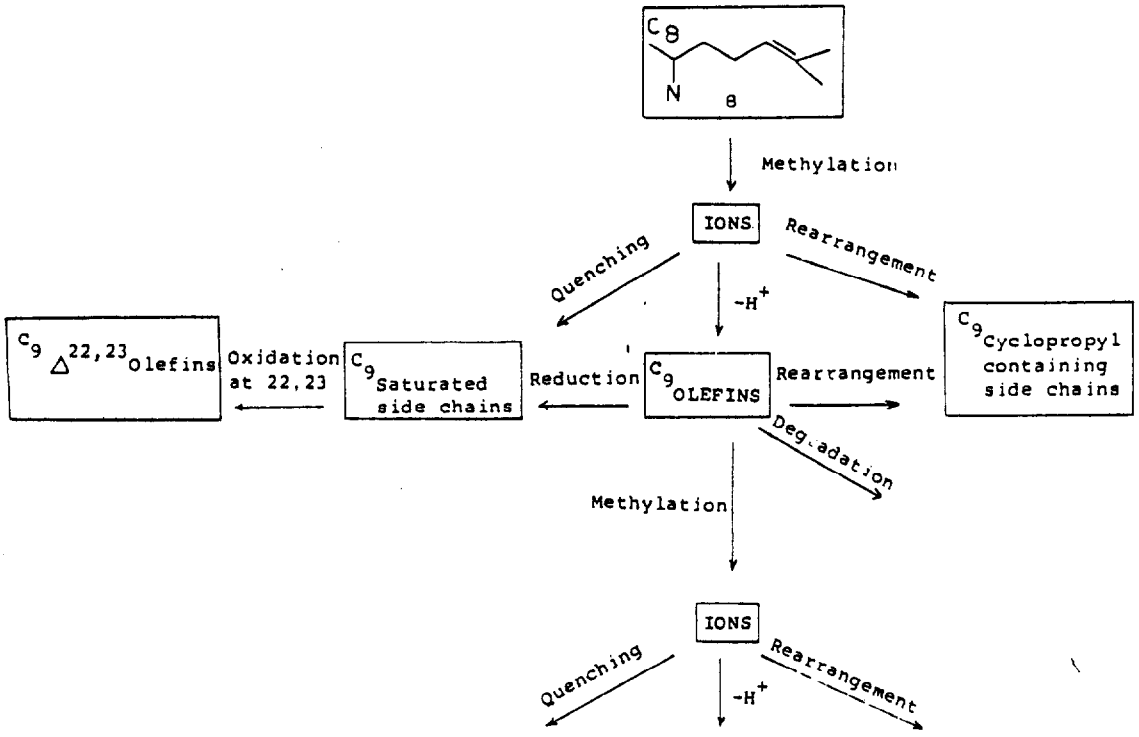
The set of reactions used in REACT to carry out possible transformations of sterol side chains have been suggested

---

[4] S. Popov, R. M. K. Carlson, A. Wegmann and C. Djerassi, Steroids 28, 699 (1976).

[5] C. Djerassi, R. M. K. Carlson, S. Popov and T. H. Varkony, in "Marine Natural Products Chemistry" by D. J. Faulkner and W. H. Fenical (ed.), Plenum : New York, N.Y., 1977, p 111.

previously [6] The precursor, (a 24,25 unsaturated side chain numbered 8 at the top of the following chart) the order of application of the various reactions and the classes of products which result are shown in the following chart. The sequence of reactions consists of repetitive application of the following steps:



[6] E. Lederer, Quart.Rev.,Chem.Soc. 23, 453 (1969).

1) Methylation. C-methylation of a double bond. In nature this reaction occurs via the ylide of S-adenosylmethionine. This reaction is constrained for general application later in the sequence to forbid the sterically unfavorable methylation of tetra-substituted double bonds.

2) The carbonium ion obtained by the alkylation can undergo several reactions:

a) proton elimination and formation of a double bond; b) cyclization to form a cyclopropyl system with subsequent elimination of a proton; c) quenching to form saturated side chains.

3) The olefin is allowed to undergo several additional reactions:

a) reduction to form a saturated side chain; b) rearrangement to a cyclopropyl system; c) degradation to shorter side chains via loss of allylic methyl groups; d) methylation to produce longer side chains.

Constraints on reactions of the olefin included

a) subsequent migration of the double bond is not allowed; b) olefins obtained by degradation are allowed to undergo only one step of methylation.

4) Subsequent oxidation of saturated side chains proceeds to form a new double bond at C-22,23, a mechanism proposed by Knapp, et al. [7] This set of reactions was applied sequentially a total of three times. Thus, side chains possessing from seven to eleven carbon atoms are accessible by this sequence.

Results. A numerical summary of results is presented in our Table below. The table is organized by summarizing the side chains produced by the different biochemical pathways. The only known, naturally occurring C7 saturated side chain was correctly predicted by REACT. Three C7 unsaturated side chains were predicted. Two of these three exist in nature. In the C8 series five unsaturated side chains out of 12 predicted are observed in nature. For the longer side chains, more are possible but fewer are observed. For example, only one out of the 76 predicted C11 side chains has so far been found in nature.

---

[7] F. F. Knapp, J. B. Greig, L. J. Goad and T. W. Goodwin, J.Chem.Soc.,Chem. Comm. 707 (1971).

Number of C in side chains	SATURATED				OLEFINS					CYCLOPROPANES		
	A	B	E	Nature	A	B	E	F	Nature	C	D	Nature
7	-	1	-	1	-	2	-	1	2	-	-	-
8	1	2	-	1	1	6	3	2	5	-	-	-
9	1	7	-	1	4	13	6	4	4	2	4	-
10	3	12	-	4	13	17	19	8	6	8	11	1
11	8	-	8	1	31	-	37	8	1	17	21	1

A methylation only.
B methylation followed by degradation only.
C rearrangement of carbonium ion.
D rearrangement of olefin.
E degradation followed by methylation only.
F oxidation of saturated side chains at 22,23 position.

Table I. Number of Side Chains Produced by Different Pathways.

The total number of sterols which obey our biosynthetic constraints is 1778. This number is manageable by techniques of computer-assisted structure elucidation. Separating the structures by molecular weight reduces considerably the number of candidate structures which must be considered in a given problem. Thus, in a GC/MS experiment the maximum number of structures we have to consider is not larger than 264 (the number of isomers with empirical formula $C_{29}H_{48}O$. Any additional spectroscopic or chemical data reduce this number still further. For other molecular weights the number of possibilities is considerably fewer. Structural information from the mass spectral fragmentation pattern of the molecule may leave only a small number of possibilities from which to choose.

2.4.1.2 Elucidation of Biosynthetic Pathways

Elucidation of biosynthetic pathways can be accomplished in several ways, including for example co-occurrence of structurally related compounds or use of mutant organisms which accumulate

intermediates. [8] These methods usually leave the structures of intermediates and/or the details of the biochemical pathways open to question. More detailed experiments are required to establish rigorously reaction pathways from precursor to product.

Isotopic labelling experiments are capable of providing additional detail through synthesis of labelled precursors followed by incorporation of labelled substrate and determination of the labelling pattern of the products of biochemical transformation. The incorporation of labelled precursors into desired products is generally low and elucidation of the labelling pattern in minute amounts of product is difficult. Thus, these experiments are generally time consuming and costly. They can be complicated by the existence of different biochemical pathways, some of which yield products with the same distribution of isotopic labels. Therefore, care must be used in designing such experiments. It is important to select a labelled precursor that will allow one to distinguish among most of the possible pathways, and that will lead to a product with labels distributed in easily detectable positions. Manual methods are often insufficient to determine all the theoretically possible pathways when the number of possible pathways and the number of intermediate structures is very large. However, this type of problem is easily managed by REACT, which can accurately and systematically monitor transformations of the precursor into products, follow the isotopic labels throughout a reaction sequence and detect the formation of equivalent structures and labelling patterns. We stress that this is not an exercise in "paper chemistry", but a systematic way to investigate all the possible aspects of a proposed experiment before devoting valuable time and resources to an experiment which leads to ambiguous results.

An example which illustrates our method is the exploration of biosynthetic pathways leading to formation of a family of fungal metabolites [9] The complete paper [22] describes our results in detail. Briefly, use of REACT enabled us to: 1) verify proposed pathways and suggest alternatives; 2) demonstrate how different patterns of isotopic labelling lead to unambiguous assignment of pathways for certain molecules; and 3) demonstrate that several pathways are possible for certain other fungal metabolites, pathways which would not be differentiated by proposed labelling schemes.

2.4.2 Applications to Structure Elucidation

The first version of REACT and its applications were

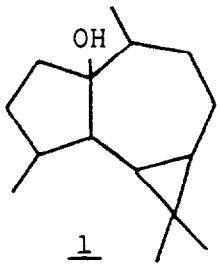
---

[8] J. D. Bu'Lock, "The Biosynthesis of Natural Products", McGraw Hill, New York, N.Y., 1965, p.94.

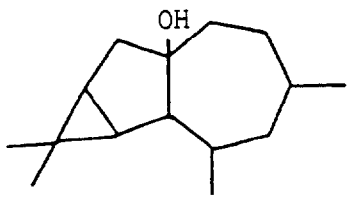
[9] G. A. Cordell, Chem. Rev. 76, 425 (1976).

described previously [22]. Subsequently, the structure of the program was revised significantly to include commands and internal operations which more closely parallel laboratory procedures. The new version has been described briefly and some applications of REACT to mechanistic problems have been discussed [24]. In subsequent sections we describe the REACT program in detail, together with an example of the application of the program to a structural problem.

To demonstrate the application of REACT we choose an example which illustrates some (but not all) aspects of the use of REACT in a structure elucidation problem. A contrived example might illustrate many of the other features and subtleties of the program, but would not be as meaningful chemically. The example involves a dehydration reaction (see reaction definition) applied during the course of elucidation of the structure of palustrol (1)[8]. Structural features of the products were powerful constraints on the identity of the compound. This problem was solved prior to the existence of the REACT program.



1



2

We pick up the example at the point at which the reaction was applied in the laboratory. This example is of interest because it represents a case where direct translation of observations on products back to structural constraints on the starting materials is difficult. Using REACT, expression of structural information is straightforward and logical. The laboratory reaction, separation and key structural information are summarized below. The starting materials, in a flask called STRUCS, are the candidate structures for palustrol (1).

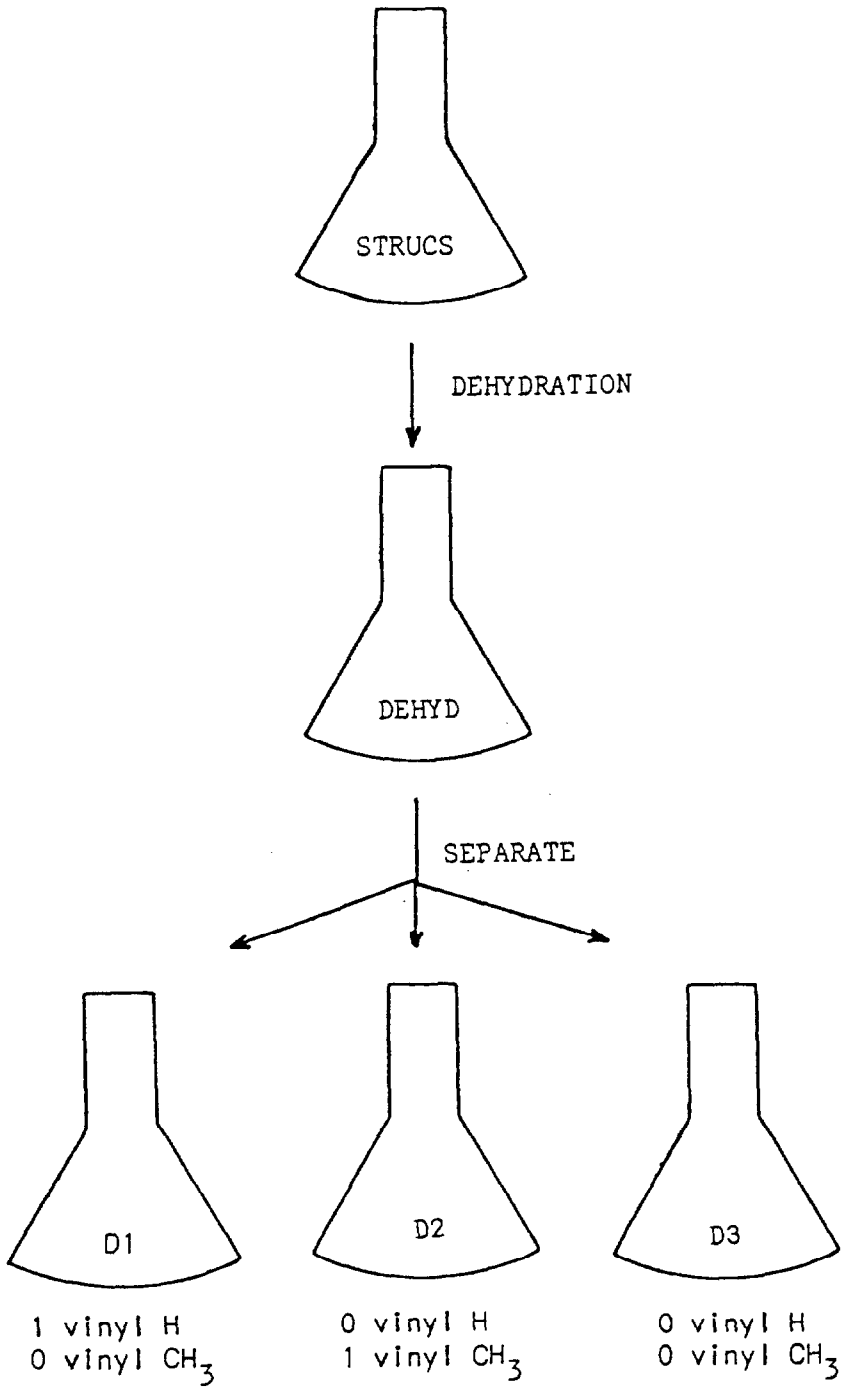


Figure 1. Diagram of REACT's Separation of CONGEN Structures with Respect to Dehydration.

Consideration of all available spectroscopic data had reduced the problem to a set of 88 candidates prior to carrying out the dehydration reaction. The contents of the flask STRUCS were dehydrated and the products placed in a flask called DEHYD. Separation of the reaction mixture yielded three products, placed in flasks D1, D2 and D3. The numbers of vinyl protons and vinyl methyl groups detected by H NMR for each product are summarized in Figure 1.

2.4.2.1 The Reaction Tree

The reaction tree is a representation of the sequence of laboratory procedures (reactions and separations) to which precursors and their products have been subjected. Formally, it consists of named flasks and their interrelationships in the form of reaction names and separation steps. If there are multiple precursors (i.e. more then one structure in a flask), as in the example, each is allowed to react, independently, resulting in a data structure internal to REACT which records the reactions of each structure separately. The chemical meaning of multiple structures in the starting material flask STRUCS is that the exact identity of the compound is not known; its structure is represented by one of all the possible structures in the flask. If the flask was created via a reaction(s), the structures represent the collection of all products from all precursors where, again, the identity of each of the products in the laboratory application of the reaction is not necessarily known. In our representation, an example of which is shown in Figure 1, flasks which could possess multiple structures, such as multiple candidates for an unknown, are depicted as containing all structures, and all possible products appear lumped together in a product flask. The dehydration reaction applied above (see Table III) is summarized in Fig. 2.

```
STRUCS=88
|
*DEHYDRATION->DEHYD=241
```

Figure 2. Result of Dehydration in REACT

This figure is interpreted to mean that the 88 candidate structures, any one of which could be the true unknown in the flask STRUCS, yield a total of 241 possible products, all associated with the flask DEHYD. Confusion related to this presentation can be avoided by remembering that the internal representation is effectively n copies of the reaction tree where n is the number of precursors in the flask STRUCS, or 88 for the

example of Fig. 1 For example, one such copy encodes the information about the conversion of 3 to 4a - 4c.

In our example we discuss only a single reaction. In general, however, the reaction tree can be of arbitrary complexity. Several different reactions can be applied to aliquots of a precursor (whether it be an original starting material or a product of a previous reaction). In addition, an extended sequence of reactions can be carried out. Thus, the reaction tree can grow arbitrarily in width and depth.

2.4.2.2 Separation

A flask obtained by reaction can contain a mixture of products. A single precursor can yield multiple products in three ways in a reaction: 1) presence of multiple reaction sites, each yielding a different product; 2) multiple reactions; and 3) cleavage reactions where all fragments are isolable. The usual laboratory step subsequent to reaction is separation of the products. Thus, REACT has a SEPARATE command which allows the chemist to express to the program his laboratory observations on performing the separation. The number of products obtained on separation is a constraint on the identity of the starting material, and is information useful in applications of REACT to structural problems. The separation requires placement of each separated product into a designated or named, flask (Table II).

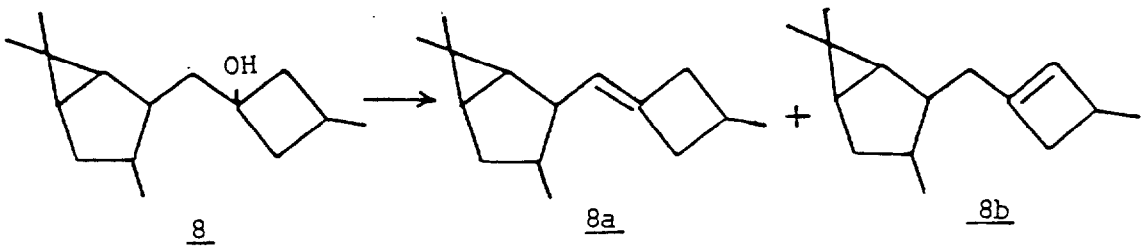
Table II. The Dialog with REACT on Separation of Contents of Product Flask
DEHYD into Flasks D1, D2 and D3

Command	Comment
SEPARATE	Enter separation mode
NAME OF FLASK TO BE SEPARATED:DEHYD	Select product flask
NEW FLASK NAME:D1 NEW FLASK NAME:D2 NEW FLASK NAME:D3	Select names for flasks for three separated products
NEW FLASK NAME:	No other flasks
MAXIMUM NUMBER OF ADDITIONAL PRODUCTS:0	No additional products in the tar flask
210 STRUCTURES SURVIVED SEPARATION	Results
BEGINNING RAMIFICATION...DONE	Implications of separation
	Return to REACT

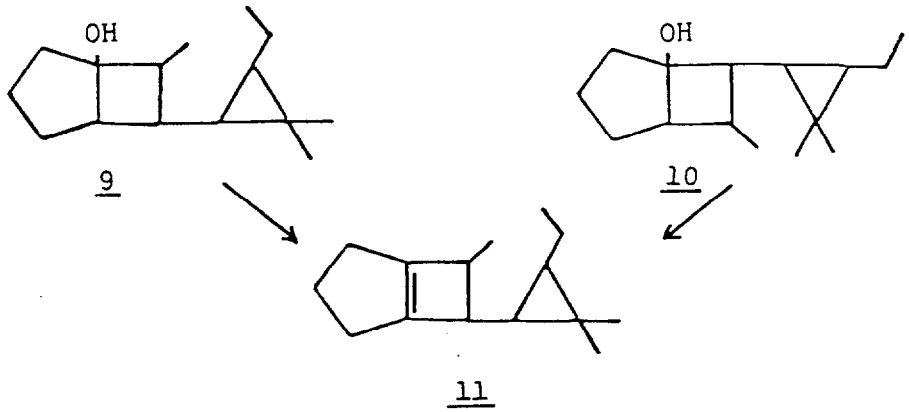
It is characteristic of many laboratory reactions that an unspecified, perhaps large, number of additional products are obtained, some legitimate, but at low concentration, others from side reactions which may not be incorporated in the definition of the reaction used in REACT. The chemist using REACT must base his use of the SEPARATE command on his own evaluation of the reaction applied in the laboratory. Selection of a named flask in which to place a separated product implies that the product so separated arose from the named reaction, and not from some other unspecified reaction. However, to accommodate the fact that the reaction may have been incomplete or side reactions may have occurred, additional products can be specified to be in a "tar" flask associated with each set of separated products. On separation, the new flasks each contain one unique product, whose identity is not known. The structure of the product must be one of the structural possibilities associated with the flask. However, the structures in the "tar" flask, (or in any flask prior to separation) can be a mixture of products, where each product in the mixture may be represented by several structural possibilities.

The dialog to establish separated products and a tar flask with REACT is summarized in Table II. In the laboratory, separation yielded three products (Fig. 1). In this example we choose to specify exactly three products by selection of three flasks to receive the products, D1, D2 and D3, and no other.

The fact that three products, all assumed to arise from the dehydration, were observed is a constraint on the identity of the starting material in the flask STRUCS. Those structural possibilities (according to CONGEN) for palustrol which would yield only two products (e.g., 8, to yield 8a and 8b) can be rejected independently of the identity of the products, while those structures which yield three products on dehydration remain under consideration (e.g., 1 and 2) until additional data on the identities of the products are gathered and specified to REACT (see subsequent section).



8 8a 8b



The reaction tree which results from the separation (Table II) is shown in Figure 3.

```
STRUCS=72
|
*DEHYDRATION->DEHYD=210-s-|D3=210
                          |
                          |D2=210
                          |
                          |D1=210
```

Figure 3. Results of Separation in REACT.

The reduction in the numbers of structures in flasks STRUCS and DEHYD (compare Fig. 3 to Fig. 2) results from the implications, or ramifications, of the statement on separation. REACT has a record of how many products are obtained from each structure and the identities of each precursor and product. It can eliminate automatically from further consideration precursors which yield an undesired number of products. If three products are observed, as in the example, only 72 of the original 88 structures remain as candidates. Sixteen of the structures yielded, by the computer program, other than exactly three products and were therefore removed from consideration as candidates. The products of these sixteen structures are also removed from the product flasks, resulting in a decrease in the number of structures in DEHYD from 241 to 210. The remaining 210 structures are not exactly three times 72 because several candidates yield equivalent products. For example, the dehydration of both 9 and 10 yields, among other products, 11.

As mentioned previously, duplicate structures are detected and removed for efficiency, except in mechanistic reactions.

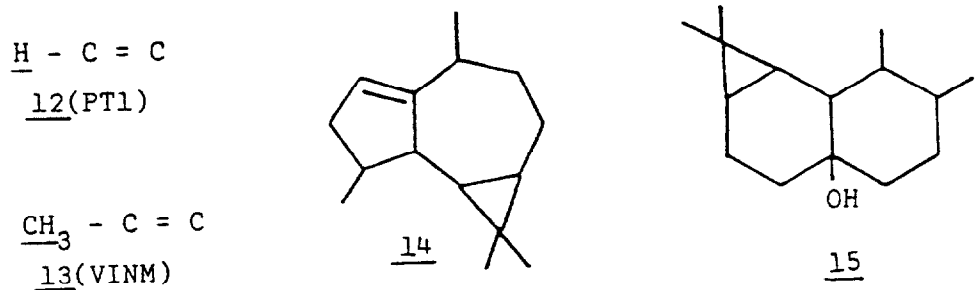
What of the contents of the flasks D1, D2 and D3? Up to this point, no statements about the structural identity of any product have been made, paralleling the laboratory events of, first, separation, and, later, gathering of data on the products. Thus, any of the 210 products in DEHYD might be in any of the flasks D1 - D3 (see Fig. 3, where all 210 products remain allocated to D1 - D3). Stated at the level of internal representation in REACT (see also above discussion), where the original structures are represented individually, each structure (in STRUCS) yielded three products, any of which might be in any flask. Subsequent operations will perform the appropriate allocations of structures to flasks.

Details of the internal representation and the algorithm which performs ramification after SEPARATE and PRUNE (see below) are given in a separate publication. This algorithm is responsible for determining legal allocations for structures to flasks throughout the reaction tree whenever the tree is modified in any way.

2.4.2.3 PRUNE - Expression of Constraints on Products

In laboratory procedures, the next step would be to collect data on the product in each flask. Structural information gained represents constraints not only on the identity of the products, but also on the identity of the precursor and its precursor and so forth throughout an entire reaction sequence. REACT allows structural statements to be made as constraints on the contents of any flask in a reaction tree. The command to express constraints is PRUNE (a word which is jargon but does carry with it the concept of trimming the reaction tree and also corresponds to the same command in CONGEN.

Substructural constraints can be obtained from a file or defined by the chemist as required, using EDITSTRUC. In our example, the product in one of the flasks (D1) was observed according to H NMR analysis to possess one vinyl proton and no vinyl methyl groups. These substructures, PT1 (12) and VINM (13), respectively, were defined and the substructures supplied to PRUNE.



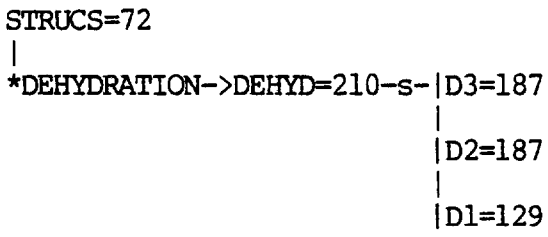


Figure 4. Application of PRUNE in REACT.

The reaction tree which results on application of PRUNE is shown in Figure 4. There remain 129 structures which could be in the flask D1. The number of structural candidates (72) has not been reduced, implying that all 72 can yield at least one structure possessing one vinyl proton and no vinyl methyls. Some candidate structures can yield more than one product which obeys these constraints and might therefore be in D1, resulting in 129 rather than only 72 structures in that flask. For example, 2 yields two products obeying the constraints; either could be the product observed in D1. However, for structure 1, only one of the products (14) is a legal structure under the constraints; that structure must be in flask D1.

If one product is forced to be in a certain flask it can be in no other flask. Thus, the number of dehydration products which could be in D2 and D3 decreases from 210 to 187 (compare Figs. 3,4). Obviously, with a more complex reaction tree, such logical decisions become complicated. REACT determines allowable allocations automatically.

Flask D2 contains a product which possesses no vinyl protons and one vinyl methyl group (Fig. 1). Constraining the contents of D2 with this structural information results in the allocation summarized in Figure 5.

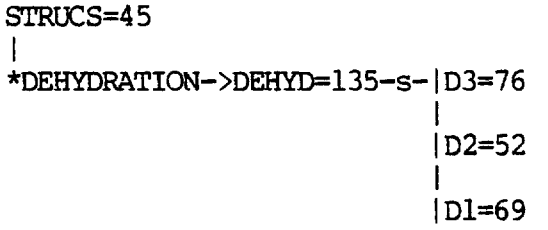


Figure 5. Results of Constraining Contents of Flasks in REACT.

Now the number of candidate structures in STRUCS is reduced to 45, implying that there are 72-45=27 structures which cannot

yield a product distribution which satisfies the structural constraints placed on both flasks D1 and D2. An example is 12, which, although it yields at least one (two) products satisfying the constraints on flask D1, yields no products satisfying the constraints on flask D2. It is therefore discarded as a candidate structure. At the same time, any products of discarded structures (and precursors in a more complex tree) are removed from DEHYD and flasks D1 - D3.

Application of the constraints on flask D3 (Fig. 1), that the product contained therein possess neither a vinyl methyl nor a vinyl proton results in the reaction tree shown in Figure 6. Now only fourteen structural candidates remain, and from the allocation of products to flasks (Fig. 6a) each yields three unique products. Each of the structural candidates was tested for the presence of exactly two secondary methyl groups; the reaction tree of Figure 7 results.

Previously, translation of the results of the dehydration into a substructure used to test the 88 candidates reduced the number of candidates to 22, rather than 14 (Fig. 6a).

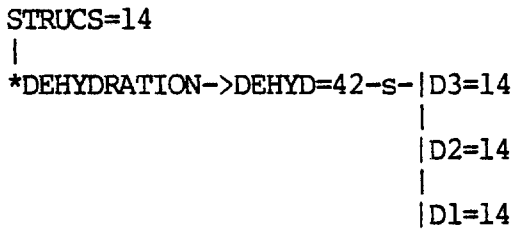


Figure 6. Further Application of Constraints.

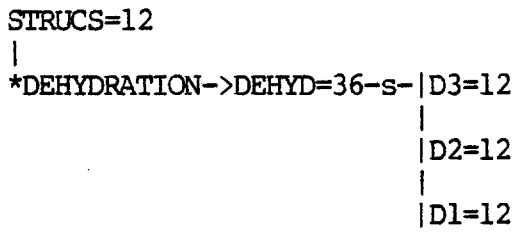


Figure 7. Constraining Contents of Flasks Still Further.

The substructure used was correct, but incomplete in that eight structures which obeyed the substructural constraint could not yield the observed products. Through use of REACT, structural information can be applied directly to the structures of potential products without the necessity of translating observations back to the precursors.

2.4.3 Utilities

We discuss the utilities briefly here not because they are critical to understanding the method but because they are an essential part of the interactive nature of REACT.

1) **Displaying Reaction Tree.** Examples of reaction trees in Figures 2-6 illustrate the format in which the reaction sequence can be observed. The DISPLAY1 command allows the chemist to view selected portions of the tree, i.e., one named flask together with any separations or reactions performed on that flask.

2) **Drawing Structures.** The structures (or any subset) in any selected flask can be drawn. To check numbering of atoms, particularly in the use of MREACT, structures can also be drawn with structure numbers (NDRAW).

3) **Determining Structural Relationships.** Relationships between precursors and products can be obtained using the PARENTS and PRODUCTS commands. A report can be obtained for all or selected structures in a flask, either to summarize precursors which led to a structure (PARENTS reports flask and structure number of every parent of every structure) or products of all or selected structures (PRODUCTS reports flask and structure number of every product of every structure). These commands were used to examine the reaction tree in the example to determine relationships among structures presented in the text.

4) **File Manipulation and Other Commands.** These utility commands allow a chemist to save and restore problems or portions thereof at will, thereby maintaining a computer-based "lab notebook" of his operations. Other commands simplify the reporting of problems and subsequent improvement of REACT and correction of errors. CHECKPOINT and UNDO are useful when the chemist wants to explore the consequences of a separation or pruning and still return to his previous reaction tree if desired.

2.5 Mass Spectral Prediction and Ranking

2.5.1 Predicting Spectra Using MSRANK and the Half-Order Theory

The MSRANK program has been incorporated as part of CONGEN, but is not yet available for general use by outside persons accessing CONGEN. We have during the past year been giving the program some extensive tests to determine its scope and limitations. We have studied the following classes of compounds (all closely related to current research problems): 1) marine sterols; 2) substituted pregnanes; 3) aliphatic and aromatic esters; and 4) macrolide antibiotics.

We conclude that MSRANK is a powerful filter for eliminating from further consideration structures which cannot yield the observed mass spectrum for an unknown by "reasonable" fragmentation pathways. The greater the structural diversity of isomeric candidates for an unknown, the better the performance of MSRANK in focussing in on the correct structure. When the structures are quite similar, for example when they have been constructed from the same set of superatoms and few remaining atoms, the ranking by MSRANK is quite similar (as one might expect). When this situation occurs, the chemist must still consider the top 10 - 50 percent of the structures as possibilities, depending on the distribution of scores.

We have added an explanation feature to MSRANK. Upon request the program prints a list of peaks in the observed spectrum which have different "reasonable" explanations. for different candidate structures. Based on this information the chemist can accept the ranking or change the parameters which define his theory of fragmentation to obtain a different ranking. This procedure helps detect and reduce the plausibility of "nonsense" fragmentation processes.

2.5.2 Prediction Using Fragmentation Rules Supplied by Chemists

When the candidate structure is known to belong to a previously investigated class of compounds, then we can use additional information to predict a more precise mass spectrum. This information is in the form of specific fragmentation rules. These rules are described by a subgraph, a break (or cleavage) and related hydrogen or neutral transfers, intensity ranges associated with rules and a parameter describing the confidence in a rule. We are working on a program which allows the user to enter rules defining his theory of mass spectral fragmentation. A computer session for entering rules which describe fragmentation of ring D in 17-substituted steroids is presented below to convey the nature of a fragmentation rule and associated parameters.

```
@<wcw>dendrl                        <begin program>
using <LISP>CARHART.SAV;70702
<WCW>DENDRL.SAV;8 created 26-JAN-78 06:06:39
what do you want to do? : CRF
create user rule file.

new rule set.
 = ?                                <query for options>
one of the following:
RESTORE  ENTER  DELETE  SHOW  SAVE  QUIT  ??
 = ENTER R1                         <enter rule named "R1">
enter rule:
 := SHOW                            <query rule>
```